

PR #29029 完整报告

sgl-project/sglang

[NPU][Bugfix] Fix a ModelSlim loading failure

合并时间: 2026-06-29 09:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29029>

执行摘要

- 一句话: 修复 ModelSlim 加载失败
- 推荐动作: 该 PR 虽小但正确, 值得合入。开发者应确保添加对应测试, 覆盖 `get_linear_scheme` 返回 `None` 的场景, 防止类似回归。

功能与动机

修复 GLM-5.2/DSA 风格模型在 Ascend NPU 上 ModelSlim 加载失败的问题。当 `get_linear_scheme()` 返回 `None` 时, `get_quant_method()` 仍返回 `ModelSlimLinearMethod`, 导致初始化时 `ModelSlimLinearMethod.create_weights()` 尝试调用 `layer.scheme.create_weights()` 时报 `AttributeError: 'NoneType' object has no attribute 'create_weights'`。

实现拆解

1. 定位问题来源: 在 `modelslim.py` 的 `get_quant_method()` 中, 对于 `LinearBase` 类型的层, 调用 `self.get_linear_scheme(layer, prefix_in_quant_config)` 后直接返回 `ModelSlimLinearMethod(self)`, 未检查 `layer.scheme` 是否为 `None`。
2. 添加空值检查: 在 `get_quant_method()` 中, 调用 `get_linear_scheme()` 之后, 立即检查 `layer.scheme is None`, 若成立则返回 `UnquantizedLinearMethod()`, 避免后续 `ModelSlimLinearMethod` 使用 `None` 的 `scheme`。
3. 保持原有逻辑: 对于有效量化方案的层, 继续使用 `ModelSlimLinearMethod`; `get_linear_scheme()` 本身不变, 仍可返回 `None` 或不支持的方案。
4. 仅修改一处: 变更只涉及 `python/sglang/srt/layers/quantization/modelslim/modelslim.py` 中的 `get_quant_method()` 方法, 新增 2 行代码, 无其他文件修改。

关键文件:

- `python/sglang/srt/layers/quantization/modelslim/modelslim.py` (模块 量化层; 类别 `source`; 类型 `data-contract`; 符号 `get_quant_method`): 核心修复文件: 在 `get_quant_method()` 中增加对 `get_linear_scheme()` 返回 `None` 的处理, 回退到 `UnquantizedLinearMethod()`。

关键符号: `ModelSlimConfig.get_quant_method`

关键源码片段

python/sglang/srt/layers/quantization/modelslim/modelslim.py

核心修复文件：在 `get_quant_method()` 中增加对 `get_linear_scheme()` 返回 `None` 的处理，回退到 `UnquantizedLinearMethod()`。

```
# python/sglang/srt/layers/quantization/modelslim/modelslim.py
# get_quant_method 方法中关键改动片段

def get_quant_method(
    self, layer: torch.nn.Module, prefix: str
) -> Optional[QuantizeMethodBase]:
    """
    根据层类型返回对应的量化方法。
    新增：若 get_linear_scheme 返回 None，则使用未量化路径。
    """

    # ... 前面的判断逻辑不变 ...
    layer.scheme = self.get_linear_scheme(layer, prefix_in_quant_config)
    # [ 新增 ] 当 get_linear_scheme 返回 None（即该层没有支持的量化方案）时，
    # 返回 UnquantizedLinearMethod，避免后续 ModelSlimLinearMethod
    # 因 layer.scheme 为 None 而调用 scheme.create_weights 失败。
    if layer.scheme is None:
        return UnquantizedLinearMethod()
    return ModelSlimLinearMethod(self)
```

评论区精华

Review 中 [gemini-code-assist\[bot\]](#) 指出一个潜在问题：开发者最初尝试在 `get_linear_scheme()` 中返回 `UnquantizedLinearMethod()` 而非 `None`，但这样会在前向传播时导致 `AttributeError`，因为 `ModelSlimLinearMethod.apply` 会调用 `scheme.apply_weights`，而 `UnquantizedLinearMethod` 没有该方法。该评论促使最终方案改为在 `get_quant_method()` 中处理 `None` 回退。此外，[TamirBaydasov](#) 在 Issue 中询问 GLM-5.2 的具体场景，因为已有对 "FLOAT" 的跳过检查，但 PR 作者未详细回复。

- 回退时机选择：在 `get_linear_scheme` 还是 `get_quant_method` 中处理 `None`?
(correctness): 在 `get_quant_method` 中检查 `layer.scheme is None` 并返回 `UnquantizedLinearMethod()`，保留 `get_linear_scheme` 返回 `None` 的行为。

风险与影响

- 风险：风险极低。仅增加了一个 `if layer.scheme is None` 的条件判断，且该路径仅在未找到支持量化方案时触发。不影响已有量化路径，不会引入回归。未添加新测试，但变更逻辑简单，手动验证即可。
- 影响：影响范围限定在 Ascend NPU 上使用 ModelSlim 量化且部分线性层未覆盖量化方案的模型（如 GLM-5.2）。修复后这些层将正确使用未量化路径，模型可正常加载。对已使用有效量化方案的层无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27204 [AMD] Implement QuarkW4A8MXFp4MoE to support amd/gpt-oss-120b-w-mxfp4-a-fp8: 同为量化方案扩展，涉及 `get_quant_method` 类似逻辑。